# #suicidal - A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter

Pradyumna Prakhar Sinha*
pradyumna_bt2k15@dtu.ac.in
Delhi Technological University

Rohan Mishra*
rohan.mishra1997@gmail.com
Delhi Technological University

Ramit Sawhney*
ramits.co@nsit.net.in
Netaji Subhash Insitute of Technology

Debanjan Mahata†
dmahata@bloomberg.net
Bloomberg

Rajiv Ratn Shah
rajivratn@iiitd.ac.in
MIDAS, IIIT-Delhi

Huan Liu
huan.liu@asu.edu
Arizona State University

## ABSTRACT

Technological advancements have led to the creation of social media platforms like Twitter, where people have started voicing their views over rarely discussed and socially stigmatizing issues. Twitter, is increasingly being used for studying psycho-linguistic phenomenon spanning from expressions of adverse drug reactions, depressions, to suicidality. In this work we focus on identifying suicidal posts from Twitter. Towards this objective we take a multipronged approach and implement different neural network models such as *sequential models* and *graph convolutional networks*, that are trained on textual content shared in Twitter, the historical tweeting activity of the users and social network formed between different users posting about suicidality. We train a stacked ensemble of classifiers representing different aspects of suicidal tweeting activity, and achieve state-of-the-art results on a new manually annotated dataset developed by us, that contains textual as well as network information of suicidal tweets. We further investigate into the trained models and perform qualitative analysis showing how historical tweeting activity and rich information embedded in the homophily networks amongst users in Twitter, aids in accurately identifying tweets expressing suicidal intent.

## CCS CONCEPTS

• **Computing methodologies** → *Learning paradigms.*

## KEYWORDS

social media mining; suicidal ideation; health informatics

**ACM Reference Format:**
Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. #suicidal - A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3358060

*Authors contributed equally to this research.
†Author participated in this research as an Adjunct Faculty at IIIT-Delhi.


*CIKM '19, November 3–7, 2019, Beijing, China*

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3358060

## 1 INTRODUCTION

Every year close to 800,000 people lose their lives to suicide and 20 times more people attempt suicide. It is also one of the leading causes of death in 15-29 year olds[1]. Suicide is a global phenomenon for which effective and evidence-based interventions can be implemented at population, sub-population and individual levels as a preventive measure for reducing suicide attempts. Suicidality is defined as any suicide-related behavior, thoughts or intent, including completing or attempting. Recent studies [18], reveal that people suffering from suicidal ideation make use of social media networks to share information about their mental health online, with many having disclosed their suicidal thoughts and plans. Therefore there is a growing interest in the research community for utilizing the signals available on social media to identify individuals who suffer from suicide ideation in an automated manner and offer them the required help and treatment.

***Motivation*** - Previous studies have observed that the risk of committing suicide has increased among youth who participate in online forums related to suicidal discussions and are susceptible to the influence of portrayals of suicide in mass media [14]. Malicious psychological contagion effects have been reported that involves social media games and challenges to coerce vulnerable teenagers for participating in self-harming activities such as the 'Blue Whale Challenge'[19]. On reviewing recent literature on online suicide intervention and prevention, it was concluded that there is a lack of existing methods on online prevention strategies and there is a need to develop effective approaches in this domain [17]. Moreover, it would also be useful to study the linguistic characteristics and social media interactions among users in a platform like Twitter, in order to understand how suicidal ideation is communicated. This motivated us to carry out the work as presented in this paper.

***Challenges*** - While solving the task of identifying suicidal ideation in social media platforms like Twitter has the potential to impact the society in a positive way, it comes with many challenges. Lack of organized efforts towards suicide surveillance, issues concerning ethics, privacy, safety and misinterpretation of information by automated tools pose serious difficulties [18]. A major challenge in monitoring online content produced on social media websites like Twitter, Facebook and Reddit, is the humongous volume of data being generated at a fast pace from varying demographic, cultural,

[1]https://www.who.int/news-room/fact-sheets/detail/suicide

linguistic and religious communities. Apart from the problem of information overload, social media websites pose challenges for automated information mining tools and techniques due to their brevity, noisiness, idiosyncratic language, unusual structure and ambiguous representation of discourse. Information extraction tasks using state-of-the-art natural language processing techniques, often give poor results when applied in such settings. Abundance of link farms, unwanted promotional posts, and nepotistic relationships between content creates additional challenges [24].

***Our Approach*** - Emergence of social media as an important vigilance and intervention platform [23, 32], has led to active research in the area of automated suicidal ideation detection using modern statistical and machine learning approaches [28, 33]. However, these techniques mostly rely on lexical and semantic features of the textual colloquial posts shared by users in different social media channels. Recent studies have shown the effectiveness of social interactions with at-risk individuals leading to on-time intervention, prevention and treatment [27]. In this work, we hypothesize that the contextual information embedded in social media engagement, historical activities of users and homophily networks formed between like-minded individuals in Twitter can lead to substantial improvements in automated identification of suicidal ideation. The major strength of such contextual information lies in the ability to look beyond linguistic cues and make predictions based on behavioral aspects and community interaction. We look into these signals throughout this work, with the help of a new manually annotated dataset (Section 5), of 34,306 tweets developed by us solely for this purpose.

We take a multipronged approach towards studying suicidality in Twitter. Not only we develop text classifiers for identifying suicidal posts (Section 4.1), but also develop a sequential neural network model taking into account the temporal nature of historical tweeting activity of users and time-variant effect of final predictions of a classifier (Section 4.2). We learn rich user and tweet representation using Graph Convolutional Networks (GCNs) from the homophily network formed between users posting suicidal tweets and discussing them (Section 4.3). Our final proposed model that achieves state-of-the-art performance on the task of identifying tweets with suicidal intent is an ensemble of the above three models combined using stacked generalization (Section 4.4). To the best of our knowledge, we are the first one to study the problem of identifying suicidal tweets by leveraging features beyond text and propose a stacked ensemble of deep learning classifiers that outperforms the current systems based solely on text and shows the effectiveness of tweeting history of users and modeling their social network interactions. Solving the problem using a multi-faceted approach was facilitated by the dataset proposed in this work that not only annotates tweets, but also shares the network of inter-user interactions amongst strongly knit groups and historical tweeting activity of users posting on suicidality.

***Our Contributions*** - Some of the main contributions that we make in this work are:

(1) Develop a new manually annotated dataset from Twitter comprising of tweets, user-specific tweeting history and social network of inter-user interactions targeted towards identifying suicidal ideation.
(2) Implement a temporally weighted sequential model for modelling tweeting activity of users that is further used for predicting suicidal tweets.
(3) Train a Graph Convolutional Network (GCN) for learning rich representation of tweets and users posting about suicidality.
(4) Achieve state-of-the-art performance using a stacked ensemble architecture for combining individual classification models trained on text, historical activity and inter-user interactions for identifying tweets expressing suicidal intent.
(5) Perform an exploratory study of the linguistic attributes and the homophily networks formed between users tweeting about suicidality
(6) Conduct a qualitative analysis to show the effectiveness of taking into account the historical tweeting activities of users and social networks between them for accurately identifying and studying suicidality in Twitter.

Next, we give a brief description of the existing works related to ours.

## 2 RELATED WORK

Various works have been recently proposed with an objective of automating the detection of social media posts expressing suicide ideation [4, 11, 33]. Cavazos-Rehg et al. [8] performed a content-based analysis on a small number of tweets related to depression in order to derive certain qualitative insights into the behavior of users displaying suicidal behavior but did not propose any method for automatically detecting such behavior. Balani and De Choudhury [3], developed a classifier to detect the level of self-disclosure in online forums such as sub-reddits, for validating the potential of harnessing signals that facilitates community moderation and support in vulnerable self-disclosing communities. While this study included post-specific features to get information related to the engagement that a post received, it did not use any information available from the social network graph. Colombo et al. [10], performed an exploratory analysis into the structure of social network graphs to gain insights into connectivity and interactions between users having suicidal intent. They found a high degree of reciprocal connectivity, and identified hubs and bridges indicating a possible contagion effect. The cues from network features were further used to improve the effectiveness of a classification system.

Seabrook et al. [35] conducted a longitudinal study of a small set of social media users to report the associations between depression severity and variability in emotions expressed by users thereby demonstrating the effectiveness of emotion dynamics of past tweets for the task of detection. Burnap et al. [7] developed a manually annotated dataset and proposed an ensemble from hand-crafted feature sets from the structure of tweets and their text-based content which showed promising results but was limited by the resources required to extract such handcrafted features in a robust manner. Guntuku et al. [16] provide an integrative review of current research in the field with regards to viablity of datasets, feature sets, learning models and performance evaluation measures while shedding light on future scope in the relatively nascent field.

The textual information available in different social media channels is often noisy, sparse and therefore may not be enough to

develop a robust predictive model solely based on it. In order to overcome the lack of predefined lexico-semantic patterns in colloquial text, methods have been used that harness other kinds of data available in these platforms, such as metadata related to the authors of the posts, and their different types of interactions leading to organically formed social networks. A number of recent studies have proposed models for constructing social media user embeddings. Amir et al. [1], generated user embeddings that captures an users' individual word usage patterns with a model similar to paragraph2vec, for the task of sarcasm detection. In this method, only users' tweets are taken into account. Yu et al. [40], proposed two neural network models, also based on paragraph2vec, to obtain users' vector representations from word representations. Since their task was recommending tweets to users, only text was considered to construct user embeddings.

Several other works have considered multiple types of features, or views. Benton et al. [4], proposed an approach based on Weighted Generalized Canonical Correlation Analysis (WGCCA), in order to turn several aspects of user information into low-dimensional vectors, including tweets and social network information.

Rich social network graphs formed by modeling user-user interactions in social media platforms can also be used as useful features in predictive models providing better representation of users' behavioral patterns. This information is extracted by representing each user as a node in a social graph and then utilizing the structure of the graph to extract features to represent the users' interaction. Mishra et al. [26], Qian et al. [30], use social graph based features and gains considerable improvement in the task of abuse detection. Yang and Eisenstein [39], use social graph embeddings to learn a gating function that is used as an attention mechanism for learning several basis models for sentiment analysis but fails to include multi-relational graphs or embeddings learned from multiple graphs to get a more informative representation of the user.

The contributions that we make in this work are different from all such previous efforts as there has been hardly any attempt to take a combined multi-faceted approach for solving the task of suicidal ideation in Twitter. In our work, we look beyond text classifiers and leverage tweeting history of users as well as their social network communication patterns. We further propose a generic framework based on stacked ensembles that learns how to combine the trained models capturing different facets of tweeting activity in Twitteroosphere and also develop a new annotated dataset that facilitates the development and implementation of such a framework.

## 3 PROBLEM DEFINITION

The main objective of this work is to identify tweets collected from Twitter that expresses suicidal intent, often referred to as suicidal ideation, and to further study them. We treat the problem of suicide ideation detection as a binary classification problem and define it as follows:

**Problem** - *Given a labeled dataset* $T = \{t_1, t_2, ..., t_n\}$ *of tweets, the objective of the task is to learn a classification/prediction function that can predict a label* $y_i$ *for a tweet* $t_i$*, where* $y_i \in$ *{suicidal intent present, suicidal intent absent}.*

As a step towards solving the above problem we seek to find answers to additional questions.

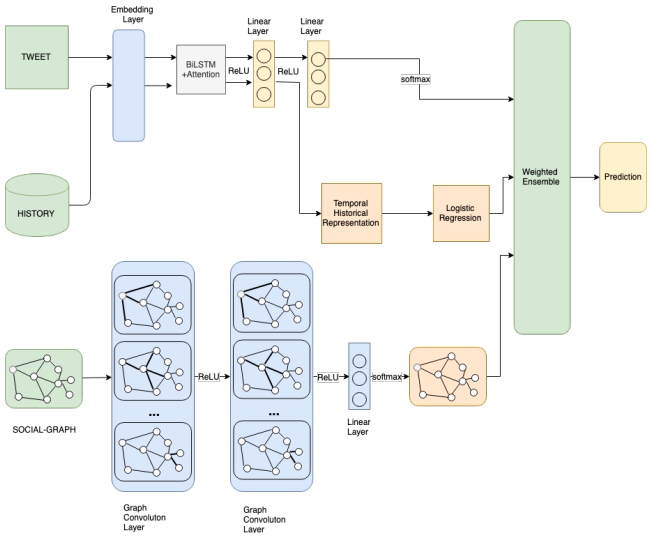


**Figure 1: Proposed architecture for identifying suicidal ideation from tweets.**

(1) Can historical activity on social media be used to supplement linguistic features for identifying suicidal tweets?
(2) Can social network interactions amongst users posting about suicidality lead to communication patterns that are helpful for identifying suicidal ideation?

We devise our methods accordingly, and describe them in details in the next section.

## 4 METHODOLOGY

We use a multipronged approach for identifying tweets related to suicidal ideation. We leverage textual as well as user-specific features for learning our models. Studies belonging to the domain of psycho-linguistics such as suicidal ideation can strongly benefit from auxiliary information that can be extracted from the metadata associated with the social media posts. They provide cues beyond that available from lexical and semantic features of text. Actions taken by an user can enable us to get a more robust understanding of their linguistic habits and affective state. User-based features were extracted from the historical tweeting activity, and inter-user interactions was modeled as a social graph. Models, were trained towards the objective of finding rich representation of users and their tweeting behavior. In order to take advantage of the diverse information encoded in the different models (text, historical tweeting activity, user interactions), we train a stacked ensemble classifier and combine the individual models using stacked generalization. The complete architecture is shown in Figure 1. We describe the details of the different components of our architecture next.

### 4.1 Text Classification Models

In order to learn from the textual information available in the raw tweets we train several text classification models and chose the best performing model as part of our final stacked ensemble. Some of the classifiers that we train on tweets are:

- **Character n-gram + Logistic Regression**: A logisitic regression model was trained taking character n-grams (ranging from 1 to 4 characters and l2-normalized) as features that works as a strong baseline [2, 26, 36].
- **GloVe + CNN**: A Convolutional Neural Network (CNN) architecture inspired from [2, 20], was trained with filter sizes (3,4,5). Each CNN layer had 100 filters and was followed by a global max-pooling layer. The embedding layer and CNN output had dropout rates of 0.25 and 0.5 respectively.
- **GloVe + LSTM**: A LSTM model with 50 LSTM units was used along with dropout layers ($p = 0.25$ and 0.5, preceding and following, respectively).
- **C-LSTM**: We replicated the C-LSTM architecture used in [33], which is currently the best system for suicide-ideation detection. It uses CNN to capture local features of phrases and RNN to capture global and temporal sentence semantics. The architecture used convolution layer with 128 filters of size 5 followed by a max-pooling layer with mask size 2. The LSTM layer comprised of 128 LSTM units with dropout rate 0.3.
- **BLSTM + Attention**: Bidirectional LSTM model with attention layers are being increasingly used to prepare sentence encodings for downstream tasks such as text classification [22, 44]. Using a Bidirectional LSTM enables capturing context over a longer span using both the left and right context of a word. By using intra-attention, the learner is able to learn the words which supply useful context with regard to the word in question. We train a BLSTM model with 100 LSTM units, dropout rate of 0.25 and a recurrent dropout rate of 0.2. The attention layer was followed by another dropout layer of 0.2. This was followed by two dense layers having 256 units and 2 units, respectively.

## 4.2 Modeling Historical Tweeting Activity

It has been shown in Brådvik et al. [6], that suicidal ideation often builds up and may be a consequence of long term depression. Therefore, learning from a representation of the user's historical tweets can extract valuable information about the affective state of the user at the time the tweet was posted.

Suppose for a user $u \in U$, the tweet under consideration is $t \in T$, $H_u$ is the set of the historical tweets and $h_i \in H_u$ is an individual historical tweet, where $U$ is the set of all users and $T$ is the set of all tweets. A BLSTM + Attention model was trained over the text-based features as described in Section 4.1. To obtain the embeddings, for each of the historical tweets $h_i \in H_u$, the latent vector was obtained from the penultimate layer of the BLSTM+Attention model. Formally,

$$g_{pt} = softmax(Dense(Dense(Attention(BLSTM(x))))) \quad (1)$$

where, $g_{pt}$ is the pre-trained BLSTM+Attention model and $x$ is an instance being fed to the model. Activation layer should be assumed after every layer. With the the penultimate layer as the output, the model is defined as,

$$g'_{pt} = Dense(Attention(BLSTM(x))) \quad (2)$$

Let, $f(h_i) \in \mathbb{R}^n$, be the latent vector representation of $h_i$, where $f$ maps the historical tweet to an n-dimensional vector representation.

$$f(h_i) = \sigma(g'_{pt}(h_i)) \quad (3)$$

In order to create a representation for the historical activity, we propose a temporal weighting scheme inspired from [43], where the importance assigned to a historical tweet varies inversely with its distance in time from the current tweet. Let $\Delta t_i$ be the time offset from the original tweet in seconds. Then, the temporal representation function $z$ is given as-

$$z(u, H) = \sum_{h_i \in H} \lambda_i f(h_i) \quad (4)$$

$$\lambda_i = \alpha \exp^{-\beta \Delta t_i} \quad (5)$$

$\lambda_i$, in Equation 4 is the weight for the temporal combination and is given by Equation 5. For each of the tweet samples, the historical activity representation of the corresponding author was considered as the feature vector and a logistic regression model was trained to learn from these features.

## 4.3 Modeling User Interactions

Learning user representations can be significantly enriched by leveraging information derived from the inter-user interactions in social media channels like Twitter, and the kind of content that users interact with. Section 5.2, describes the social graphs that were constructed to represent these interactions. Much work has been done on generating node embeddings from graphs [15][29]. Such methods are capable of extracting structural information from the graphs to assign similar vector representations to nodes that have similar roles or belong to the same network community. However, these methods are unsupervised and only look at assigning relative representations to the nodes. Only the structural information from the graph gets encoded. Graph Neural Networks, on the other hand, generate representations for the nodes by taking into account both the topological structure of the graph and the individual features of the nodes. In our work, we apply Graph Convolutional Networks for learning the representation of the nodes, as explained next.

*4.3.1 Graph Convolutional Networks.* Kipf and Welling [21], introduce an architecture for semi-supervised learning on graphs referred to as Graph Convolutional Networks (GCNs). GCNs are capable of propagating features of nodes in the graph and allow contextual learning of information with respect to a node's neighbourhood. We employ GCNs to capture the contextual linguistic information for users for an effective user profiling.

The layers of a GCN combine the input feature vector of each node with those of its 1-hop neighbourhood nodes, followed by a linear combination through a dense layer. After every layer, the neighbourhood being looked at gets widened by 1 hop. Let $G(V, E)$ ($|V| = n$) be our graph under consideration with the adjacency matrix $A \in I\!R^{nxn}$. Let $d_l$ denote the length of the output vector of the $l^{th}$ layer. With the feature matrix for the nodes $F \in I\!R^{nxd_0}$, and the output of the layer $l$, $H^l \in I\!R^{nxd_l}$,

$$H^{(l+1)} = f_l(H^{(l)}, A) \quad (6)$$

where, $f_l : \mathbb{R}^{nxd_l} \rightarrow \mathbb{R}^{nxd_{l+1}}$. The function $f_l$ gives the transformation at the $l^{th}$ layer as,

$$\hat{A} = A + I \tag{7}$$

$$\tilde{A} = \hat{D}^{\frac{-1}{2}} \hat{A} \hat{D}^{\frac{-1}{2}} \tag{8}$$

$$f_l(H^{(l)}, A) = \sigma(\tilde{A} H^{(l)} W^{(l)}) \tag{9}$$

Equations 7 and 8, defines the transformation to obtain the normalized adjacency matrix $\tilde{A}$. Equation 7 adds self loops to the graph to take into account a node's own feature vector. $\hat{D}$ in Equation 8, is the diagonal degree matrix, where $\hat{D_{i,i}} = \sum_{j \in V} A_{i,j}$. Essentially, Equation 8 normalizes the weights for each node pair as,

$$\tilde{A}_{i,j} = \frac{\hat{A}_{i,j}}{\sqrt{\hat{D_{i,i}} * \hat{D_{j,j}}}} \tag{10}$$

In Equation 9, the dot product of $\tilde{A}$ with $H^{(l)}$ does a weighted combination of its feature vector with those of its 1-hop neighbours. These vectors are then passed to $d_{l+1}$ hidden units similar to a dense layer. $W^{(l)} \in \mathbb{R}^{d_l x d_{l+1}}$, is the weight matrix which gives the linear combination for each of the neurons in the layer $l$.

In order to effectively capture contextual, linguistic, and structural information of the users, we performed semi-supervised learning on our graph using GCNs. For this purpose, an extended graph was constructed from the combined graph (as described in Section 5.2, and Table 2). The extended graph comprised of nodes corresponding to both the users (authors of tweets) and tweets. Two user nodes were connected if they shared an edge in the original graph. Each tweet node shared an edge with its corresponding user (author) node. The resulting extended graph had 66,864 nodes, with 32,558 users; 34,306 tweets; and 92,443 edges.

For the feature vectors of the tweet nodes, we extracted Tf-IDF features from unigrams and bigrams of the corpus and selected the best 1000 features using ANOVA method. For the user nodes, the vector representation was taken as the mean of feature vectors of the corresponding tweet nodes. All the feature vectors were l2-normalized. The tweet nodes were labeled while the author nodes were unlabeled. Semi-supervised learning was performed. A three-layer architecture was used - **<graph conv layer (GC-1), graph conv layer (GC-2), dense layer>**. The two graph convolution layers had 500 and 100 units respectively, and the dense layer had 2 units over which a softmax distribution was constructed for the two target classes. This architecture was chosen because of the enormity (large number of nodes and edges) of the graph. The added *Dense* layer as compared to the architecture used in [21], allows for better incorporation of a large amount of information available in the graph. The output probability vector is given by,

$$O = softmax((relu(\tilde{A}(relu(\tilde{A}FW^{(1)}))W^{(2)}))W^{(3)}) \tag{11}$$

Here, $W^{(1)}$ and $W^{(2)}$ are the weight matrices for the GC-1 and GC-2 respectively. $W^{(3)}$ is the weight matrix for the dense layer. Each layer had a dropout of *0.4*.

## 4.4 Stacked Ensemble Model

Often, one solution to a complex problem does not fit all scenarios. Thus, researchers use ensemble techniques to address such problems. Ensemble learning has proved to be very effective in most of the machine learning tasks [13, 23]. Ensemble models can offer diversity over model architectures, training data splits or random initialization of the same model or model architectures. Some of the popular ensemble approaches are bagging and boosting [31]. Stacked ensembles are also one of the widely used ensemble techniques that we leverage in this work. Stacked ensembles find the optimal combination of a collection of prediction algorithms using a process called stacking or stacked generalization [37]. It uses the predictions of the basis models (level-0 models) to train a new model (level-1 model) and make the final predictions. For the level-1 model, the input vector for an instance $x$ is comprised of the predictions from the level-0 models that were trained on a subset of the data which did not include $x$. We describe this technique formally for the k-fold cross validation split strategy next.

Formally, for a dataset $D = \{(x_i, y_i); i = 1...n\}$ where $x_i$ is the sample and $y_i$ is the corresponding label, let the splits for K-fold cross validation be $\{D_1...D_k\}$. $D_j$ and $D^{(-j)} = D - D_j$ are the test and training set respectively for the $j^{th}$ fold. Given $L$ learning algorithms, the $k^{th}$ algorithm is used to build the model $M_k^{(-j)}$ over the data in $D^{-j}$. These learning algorithms are called level-0 generalizers. For, $x_i \in D_j$, let the output probability vector of the model $M_k^{(-j)}$ be $z_{ki}$. For $c$ classes, $z_{ki} \in \mathbb{R}^c$. Stacked generalization gathers these output vectors to create a new level-1 data given as,

$$D_{CV} = \{(z_{1i}...z_{ki}...z_{Li}; y_i); i = 1...n\} \tag{12}$$

This new dataset is further fed to another model, called the level-1 generalizer, which learns how to combine the outputs from the basis models. Instead, we apply weighted averaging to obtain the final class probability vector. Let $w_k$ be the weight for the $k^{th}$ basis model with the condition $\sum_k^L w_k = 1$. The final probability vector for the instance $x_i$ is given by,

$$z_i = \sum_k^L w_k * z_{li} \tag{13}$$

To determine the optimal weights for the combination, a grid search was conducted over all the possible choices of weights with a granularity of 0.01, as described in Section 6. Next, we present a detailed description of the dataset on which we conduct our studies.

# 5 DATASET

We create a manually annotated dataset that is described in this section. The dataset generation was done in two phases: (i) A lexicon of suicidal phrases was generated. (ii) Real-time and historical tweets were collected using the lexicon along with the social engagement data associated with them. We would like to make the anonymized version of this dataset and the associated lexicon available to the research community on request.

## 5.1 Developing a Lexicon of Suicidal Phrases

In order to collect and analyze suicidal ideation in Twitter, a set of terms that were likely to identify suicidal communication were identified. Three different forums[2][3][4] that are dedicated to the discussion of suicidal thoughts were scraped and the top posts (based

[2]https://www.reddit.com/r/suicidalthoughts
[3]https://www.reddit.com/r/SuicideWatch
[4]https://www.takethislife.com

on votes) were used to extract phrases. Posts with little or no suicidal ideation information were removed manually. The TextRank algorithm [25], as implemented in Pytextrank[5], was used for ranking and gathering the list of the most prominent phrases from these posts. The resulting list had 207 phrases such as *think suicide, wanting to die, suicide times, last day, alternate life, time to go, hate life.* Furthermore, the lexicon was extended by using the lexicon shared in [33], resulting in a lexicon of 248 phrases.

## 5.2 Data Collection

*5.2.1 Tweet Collection.* For each phrase in the curated lexicon, tweets were collected using the Twitter REST API[6]. A total of 48,887 tweets were obtained. Furthermore, retweets and non-English language tweets were removed. Certain phrases such as *last day*, from the lexicon lead to tweets such as *Last day at work, holidays!*, as the data was collected during December, 2018. A manual check was done to remove such tweets (around 3,000), that were trivially non-suicidal or exhibited flippant behaviour. Moreover, 12 user accounts were detected to be throwaway accounts made within 10 days of scraping and were therefore discarded. The resulting dataset consists of 34,306 tweets. Table 1, shows certain linguistic and textual

| Attribute | Value |
|---|---|
| Avg Tokens | 30.99 |
| Avg Slang Words | 1.476 |
| Avg Length of Tweet (in chars) | 133.37 |
| Avg No. of Hashtags | 0.199 |
| Percentage of Tweets with URL | 22.58 |
| Avg no. of feeling words | 2.404 |
| Avg no. of Nouns | 10.57 |
| Avg no. of Verbs | 6.34 |
| Avg no. of Adjectives | 3.11 |

**Table 1: Statistics of linguistic and textual attributes extracted from the collected dataset.**

attributes of the dataset, which presents a high-level insight. We took the slang words and feeling words shared in [24]. This list is consolidated from resources that are developed for compiling the jargon and the language used in social media. An interesting thing to note is that the tweets, on average contain two feeling words, which is quite intuitive as posts having signals of suicidal intent will most likely have words expressing feelings of the author. Some of the top feeling words used were - *feel*, *dead*, *shit*, *wish*, *tired*, *suicidal*, *depressed*, *hurt*, *mean*. We observe very little use of hashtags in the tweets (with an avg of 0.199), and only around 22% of them share URLs. Slang words like *f**k*, *f**ked*, *c*nt*, *ass*, *b*tch*, were also frequently used. We also perform POS tagging of the tweets using the penntreebank tagger implemented in NLTK. We report the usage of different tags in Table 1.

*5.2.2 User Historic Activity Data.* For the 34,306 tweets in the dataset, there are 32,558 unique users. For each of these users, the tweet timeline (previous 3,200 tweets or as many as available) was also collected using the Twitter API.

[5]https://pypi.org/project/pytextrank/
[6]https://developer.twitter.com/en/docs.html

| Graph | Edge Representation | Avg Degree | Density ($10^{-5}$) |
|---|---|---|---|
| **quotes** ($G_{quotes}$) | A quoted B | 0.816 | 2.508 |
| **mentions** ($G_{mentions}$) | A mentioned B | 2.87 | 8.842 |
| **repliedTo** ($G_{repliedTo}$) | A replied to B | 1.917 | 5.888 |
| **Combined** ($G_{combined}$) | Union of the above three relationships | 3.571 | 10.969 |

**Table 2: Details for the social graph created from different user interactions (A and B represent users along an edge).**

*5.2.3 Social Graph Data.* In this work, we also investigate into the social graph constructed from the user interactions in our dataset (Section 4.3). Let $U = \{u_1, u_2, ..., u_m\}$, be the set of $m$ users who authored tweets in the collected dataset $T$. Four different graphs were constructed from this data with the set $V = \{v_1, v_2, ..., v_m\}$ of $m$ vertices, that are fixed and has a one-to-one mapping with $U$.

(1) *quotes graph* - A directed graph $G_{quotes} = (V, E)$, where $E = \{e_1, e_2, ..., e_p\}$ is the set of $p$ edges, where $e_i = (v_x, v_y)$ represents an edge-relationship between $v_x \in V$, and $v_y \in V$, whenever $v_x$ quotes a tweet $t_i \in T$, posted by $v_y$.
(2) *mentions graph* - A directed graph $G_{mentions} = (V, E)$, where $E = \{e_1, e_2, ..., e_q\}$ is the set of $q$ edges, where $e_i = (v_x, v_y)$ represents a directed edge-relationship between $v_x \in V$, and $v_y \in V$, whenever $v_x$ mentions $v_y$ in a tweet $t_i \in T$.
(3) *repliedTo graph* - A directed graph $G_{repliedTo} = (V, E)$, where $E = \{e_1, e_2, ..., e_r\}$ is the set of $r$ edges, where $e_i = (v_x, v_y)$ represents an edge-relationship between $v_x \in V$, and $v_y \in V$, whenever $v_x$ replies to $v_y$, by posting a tweet $t_i \in T$.
(4) *Combined Graph* - An undirected graph $G_{combined} = (V, E)$, where $E = \{e_1, e_2, ..., e_z\}$ is the set of $z$ edges, where $e_i = (v_x, v_y)$ represents an edge-relationship between $v_x \in V$, and $v_y \in V$, whenever $e_i$ satisfies any edge relationship as defined in $G_{quotes}, G_{mentions}, G_{repliedTo}$. This is basically a union ($G_{quotes} \cup G_{mentions} \cup G_{repliedTo}$) of the three graphs defined above.

Although, the set of users $V$ (vertices of the graph $G$), were fixed, the set of edges $E$, were defined between them depending on the type of interactions. Table 2, shows the different graphs constructed corresponding to four different forms of interactions and their characteristics in terms of *avg degree*, and *density* [41]. Table 3, shows examples for each form of interaction.

## 5.3 Annotation Details

Our annotation guidelines were based on the following classification system -

(1) **Suicidal Intent Present**
- Posts where suicide plan and/or previous attempts are discussed. eg *its been a long time... and i feel like i want to cut my wrist again – <url >*
- Text conveys a serious display of suicidal ideation. eg. *tw// self doubt, suicide. Im feeling so worthless right now. Im having a bit of suicidal thoughts again. The time when i thought they were gone for good....*
- Posts where suicide risk is not conditional unless some event is a clear risk factor eg:depression, bullying, etc.
- Words used are somber and not flippant.

| | | | |
|---|---|---|---|
| 1 | I came a long way from wanting to commit suicide | **quoted** | Last semester i took "Statics"course and got a "D"grade. I don't know what i was doing at the time but I am sure i was wrong. Everything is different now... |
| 2 | Remember the simpler times when life was a bitch? Now life is f**ked up | **mentioned** | • i don't know mannn. it's all f**ked up. life is f**ked up. love is f**ked up ... and then even death is f**ked up.<br>• it just f**ked me up even more. lol at life. |
| 3 | i want to die every single moment that i f**king breathe. | **repliedTo** | i wish i never had to wake up to this life again, my life is so f**king pathetic. |

**Table 3: Examples of different types of interactions in Twitter as captured by the social graphs.**

(2) **Suicidal Intent Absent**

- Posts emphasizing on suicide related news or information. eg. *Man found dead in Hong Kong hotel room with semi-conscious woman, in suspected suicide pact <url><url>*
- Posts containing no reasonable evidence that the risk of suicide is present; includes posts containing song lyrics, etc. eg. *man i've thought about suicide a hundred times but i'd hate to disappoint and see my mama cry* (Song Lyrics).
- Condolences and awareness posts, eg. *"Norfolk County Council free app #SuicidePrevention #Support <url>*

Two annotators, who are students of clinical psychology and regular social media users, were provided with the guidelines to label the tweets as used in [34]. The inter-annotator agreement was calculated to be $\kappa = 0.72$ (Cohens' Kappa). The annotation process was supervised by a practising clinical psychologist. In cases of ambiguity the default class '0 (non suicidal)', was assigned by the annotators. 1,028 of the 9,605 conflicts were resolved by discussions among the annotators and supervisor. In case of failure to resolve, a majority vote as decided by the supervisor was taken as the final label. Following the first annotation pass, the annotations were reviewed by the supervisor and in a few cases (124), the given annotation was altered following discussions. The resulting dataset had 3,984 suicidal tweets (11.61% of the entire dataset). A significant fraction of the dataset included tweets recognised to be flippant references to suicide, messages of support and awareness, suicide related news and information, which although being non-suicidal were collected by the keyword filtering approach. Next, we describe the experiments that we performed on the dataset applying the methods as explained in Section 4.

## 6 EXPERIMENTS AND RESULTS

For all our experiments we performed 10-fold stratified cross-validation, and all the experiments were conducted and evaluated on each of the 10 train-val splits. The performance metrics (F1 score, Average Precision score, Precision and Recall) have been reported as average across the 10 validation sets. The hyper-parameters for each of the models were tuned using a grid search over the validation sets.

Zhang and Luo [42] describe the lacunae of reporting metrics such as micro F1, Precision or Recall provided in cases of highly imbalanced datasets such as abuse detection. The results therefore include the weighted F1 Score, area under the Precision-Recall Curve, Precision Score and Recall Score. Statistical significance of results was determined by employing Wilcoxon Signed Rank Test [38], using ten stratified subsets from the test set with each metric chosen as a skill estimate for the different competing systems. Wilcoxon Signed Rank Test was used in favor of paired t-test as it is a non-parametric test with no assumption about the inherent distribution of the results even though it has less statistical power than paired t-test, although more power when the expectations of the t-test are violated, such as independence [12].

| Model | F1 | AP | P | R |
|---|---|---|---|---|
| **Char ngram+ LR** | 85.62 | 60.65 | 90.40 | 83.38 |
| **LSTM** | 90.60 | 64.04 | 90.47 | 91.38 |
| **CNN** | 91.67 | 71.01 | 91.61 | 92.28* |
| **CLSTM** | 91.76 | 71.32 | 91.23 | 91.67 |
| **BiLSTM+Att** | 92.26* | 73.77* | 92.20 | 92.52* |
| **Text+History+Graph** | **92.76*** | **76.10*** | **92.76*** | **92.87*** |

**Table 4: Results with weighted F1 and Average Precision Score. * indicates that the result is significantly better than CLSTM (p <0.05).**

In order to conduct the experiments, the dataset was loaded to the pipeline the first step of which is the preprocessing of tweets. The raw text of the tweet were parsed using a tweet tokenizer[7] that replaced every username mentions, hashtags, and URLs with <mention>, <hashtag>and <url>respectively. Punctuations were removed and contractions were fixed. The WordNet Lemmatizer provided by nltk[5] for lemmatization.

Following it, the data is split for cross-validation and fed into the different models i.e., the text classifiers, graph convolutional networks and the model trained on historical tweeting activity. For the deep learning based text classifiers as described in Section 4.1, the preprocessed text was encoded onto a padded sequence and used as an input for the learners. The best performing model (BiLSTM + Attention) was further tuned for number of LSTM units over the grid $\{50, 100, 200\}$ and $\{NoAttention, Attention\}$. Based on the performance 100 units were fixed and attention mechanism is used in the final text learning model. The deep learning text classifiers were trained using the Adam optimizer, an early stopping criterion with a patience of 10 epochs and a learning rate of *0.001*. All the models were implemented using Keras[8] with Tensorflow[9] backend.

For the model trained on historical tweeting activity as explained in Section 4.2, historical representations for all the historical tweets of users are inferred and used as an input for the temporal weighted combination. The hyperparameters for the temporal weighted combination that takes into account the historical tweeting activity were tuned using a grid search over the grid $\alpha = \{0.1, 0.5, 1.0\}$ and $\beta = \{0, 0.01, 0.1, 1\}$, yielding $\alpha = 0.5$ and $\beta = 1$. The results shown in the Table 5 compares the performance of temporal and non-temporal weighting scheme.

[7] https://pypi.org/project/tweet-preprocessor/
[8] https://keras.io
[9] https://www.tensorflow.org

The preprocessed text and combined graph were passed as inputs into the social graph model as explained in Section 4.3. The hyper-parameters for the size of feature vectors to represent the nodes was conducted over the grid $\{1,000; 2,000; 5,000\}$. The Graph Convolution Network was trained with Adam optimizer, setting the learning rate to 0.01, a dropout rate of 0.4 and an early stopping criterion with a patience of 10 epochs. The implementation was done in Tensorflow. For all our future descriptions we refer the text classification model by **Text**, the model trained on historical tweeting activity as **History**, the GCN model trained on the social graph as **Graph**, and whenever we combine them with the ensemble model we denote the combination using the + symbol.The results of the text classifiers are reported in Table 4.

The results of the text classifiers are reported in Table 4.

| Strategy | F1 | AP |
|---|---|---|
| Non-Temporal | 0.8426 | 0.2614 |
| Temporal | 0.8904 | 0.5005 |

**Table 5: Performances of the temporal and non-temporal weighting scheme for modeling historical tweeting activity.**

To create the weighted ensemble as described in Section 4.4, a grid search was carried for the optimal weights. The search space comprised of all possible weights to a granularity *0.01*. For the **Text+History** combination, the optimal weights were 0.84 and 0.16 and for the **Text+History+Graph**, the optimal weights were 0.52, 0.10 and 0.38. In Table 5, it can be observed that temporal modelling of user tweeting history ( as in **History** ) led to an effective classifier even without relying on the text of the tweet. It showed significant improvement over the baseline of non-temporal modelling of historical tweet representations. In Table 6, it is interesting to observe the ability of **History** and **Graph** models to classify the tweet without relying on the textual features to a sufficient degree. When combined with the **Text** model, the **History** model shows a minor gain in performance. However, by combining all the models, a significant gain is observed.

| Model | F1 | AP | P | R |
|---|---|---|---|---|
| Text | 92.26 | 73.77 | 92.20 | 92.52 |
| History | 89.53 | 57.84 | 90.40 | 91.21 |
| Graph | 90.29 | 71.47 | 91.73 | 89.54 |
| Text+History | 92.33 | 74.03 | 92.26 | 92.60 |
| Text+History+Graph | 92.76 | 76.10 | 92.76 | 92.87 |

**Table 6: Ablation study of the different model combinations.**

# 7 FURTHER EXPLORATION

## 7.1 Qualitative Analysis of Model Predictions

We perform a qualitative ablation study and error analysis by hand-picking some peculiar data instances in our dataset, selecting them from the true positive and false positive results of the different models trained in this work. These results are shown in the Table 7. In examples 1 and 2 of the Table 7, the information available from the text is sparse and does not contain any explicit indicators for suicidal ideation. Therefore, the **Text** model fails to classify the text correctly. However, the author in example 1 has written multiple other depression related tweets. In example 2, the user's profile was checked and there were limited tweets available which made the information available from the history sparse. However, the user was in close proximity to other suicidal users and therefore our final ensemble model makes an accurate classification.

It is often the case that tweets indicative of suicidal ideation are highly subjective in content with multiple, often contradictory, phrases. Therefore, classifiers that look into the text-only features sometimes fail to understand the pragmatic meaning of the given document. In examples 3, 4 and 5 of Table 7, the textual content is long and noisy making the task of classification non-trivial for the text only classifier. However, our final ensemble model relies on the position of the users in the social graph along with historical activity and is able to correctly detect suicidal ideation in the tweet.

Other than being able to better detect tweets related to suicidal ideation, we also observed a marked decrease in false positives i.e. tweets that were inaccurately detected to be suicidal. In example 6 of Table 7, though there is an explicit mention (*i'll kill myself*) it is understood to be a frivolous remark. By leveraging on the information from the historical data our ensemble models accurately classify the tweet to be non suicidal. Moreover, examples 7, 8 and 9 of Table 7 are lengthy and the pragmatic meaning of the tweet are known to be messages of awareness and support. Even though there was a history of tweets with negative emotional connotation in the user profile of example 8, the current tweet is indicative of successful recovery and our ensemble model that uses graph based features correctly classifies the tweet.

A limitation of current approaches of suicide ideation detection from text-based features is the inability to discern movie quotes and song lyrics from original writing. Considering example 10 and 11 of Table 7, it can be observed that text-only models as well as historical features, fail to identify that the tweets contain song lyrics. However, the information extracted from the social graph is a decisive factor in correctly classifying them as non-suicidal thereby eliminating them as false positives.

However, some cases were encountered during the ablation study where our ensemble failed to accurately classify the tweet. In example 12 of the Table 7, the tweeting history of the account was sparse and the user was not connected to any other users in our dataset. Therefore, no contextual information was available for the user leading to the misclassification.

## 7.2 Analyzing the Social Graph

With the aim of understanding the communication and propagation of suicidal content, we perform a social network analysis of the combined graph $G_{combined}$. For this purpose, we take the subgraph containing only suicidal users and eliminate the solitary nodes. This sub-graph was then analyzed on the criteria of various metrics to get a better idea of the level of association among the users. It was done to understand whether there exists closely knit networks among suicidal users or do the users exist in smaller groups and are isolated from each other. The results are shown in Table 8. It can be observed that there does exist a significantly well connected group of suicidal users with a potential for studying the spread of information amongst them. It could be possible that users rely on

| ID | Text | Label | T | TH | THG |
|---|---|---|---|---|---|
| 1 | End my life | 1 | 0 | 1 | 1 |
| 2 | Don't want to be here. | 1 | 0 | 0 | 1 |
| 3 | I'm far from beating my demons but tonight has been a particularly rough night for me. .<br>It reminded me of one other night when I was rock bottom and wrote out a suicide note.. | 1 | 0 | 0 | 1 |
| 4 | time I do anything to try to make myself happy it makes them even more fucking depressed.<br>I honestly wish something would happen out of my control like a car accident or getting shot or<br>something so I could diebut not have been the one to take my own life. I don't want to die | 1 | 0 | 0 | 1 |
| 5 | I wanted to live, I just didn't know how to anymore. And I think that's something<br><url>that a lot of people miss, they think that individuals who are suicidal want to die. | 1 | 0 | 0 | 1 |
| 6 | josh is so cute i'll kill myself <url> | 0 | 1 | 0 | 0 |
| 7 | lead to suicide because i lost my brother to suicide out of absolutely nowhere and there were<br>no clues as to what pushed him over the edge. You don't know the impact the s*it you say has. You don't<br>realize how f*cking hostile you can be, it's disgusting. Think before you speak | 0 | 1 | 0 | 0 |
| 8 | When I feel down I convice myself that no one likes me, wants to be around me,<br>or talk to me. Today so many people made me realize that that's absolutely not the case.<br>I'm so fucking blessed to be so loved and supported. | 0 | 1 | 1 | 0 |
| 9 | Depression & Suicide is so sad. Did you know ketamine infusion therapy is the fastest acting,<br>most effective medicine for depression, suicidal ideation, PTSD & some other mood disorders?<br>In Colorado, call 720-***-**** for a free consultation today. <url> | 0 | 1 | 0 | 0 |
| 10 | I don't wanna die But sometimes wish I've never been born at all | 0 | 1 | 0 | 0 |
| 11 | when you said your last goodbye, i die a little bit inside, i lay in<br>tears in bed all night, alone without you by my side | 0 | 1 | 0 | 0 |
| 12 | Nobody can be surprised when I snap and take someone's life.<br>Either my own or these pricks. None of you can be surprised. | 1 | 0 | 0 | 0 |

**Table 7: Examples for the qualitative abalation analysis of the trained models. T-Text , TH - Text + History , THG - Text + History + Graph**

| Metric | | Measure | Inference |
|---|---|---|---|
| Radius | Max | 6 | Short radius and diameter indicate tightness in the community connectivity and high degree of influence amongst the suicidal users |
| | Avg | 1.17 | |
| Diameter | Max | 11 | |
| | Avg | 1.62 | |
| Connected Components | Count | 138 | Suicidal users tend to cluster together through actions of quoting, mentioning and replying |
| | Max-Size | 73 | |
| | Avg-Size | 3.68 | |
| Harmonic Centrality | Max | 49.84 | High centrality measure indicates the behaviour of suicidal users as major social hubs |
| | Avg | 6.94 | |

**Table 8: Analysis of the Social Graph**

virtual relationships formed over social networks for support and solidarity.

However, it is also of paramount importance to investigate the cascading of information in the wider community where it could pose a risk for suicide contagion. To this end, the presence of hubs with higher social capital must be detected. These specific users are influential in the propagation of suicidal information and can be crucial in either extending support to suicidal users or curbing their influence by monitoring the communication made by such users to avoid a contagion effect.

## 8 ETHICAL CONSIDERATIONS AND LIMITATIONS

Analyzing social networks for suicidal ideation necessitates the need to look into challenges and limitations we face while keeping ethical considerations in mind. Based on the issues highlighted in [9] we address these as:

- **Privacy**: Individual consent from users was not sought as the data was publicly available and attempts to contact the author for research participation could be deemed coercive and may change user behavior that might disturb the integrity of the experiments conducted. We are cognizant of the ethical limitations that occur in the absence of consent and feedback from the study population.
- **Bias**: The authors are aware of the inherent biases that come with collecting data from social media which are not 'naturally occurring' but are influenced by agents that have varying sets of priorities. While using a specific set of seed keywords to build a lexicon and scrape data introduces bias, we're able to extract a diverse dataset rich in non concerning suicide-related, which is the primary resource to learn the subtleties and intricacies of suicide ideation.
- **Interpretation**: Despite a comprehensive annotation process, we acknowledge that it is likely that our interpretation of subtle nuances in suicidal ideation may be incorrect given the highly subjective nature of this challenging problem. The motive of this study was to determine if community contextual features can identify such linguistic patterns, rather than focus on the intrinsic labeling of tweets.

## 9 CONCLUSION AND FUTURE WORK

In spite of the importance of suicidal ideation identification on social media, little knowledge management research has focused on looking beyond linguistic patterns, or the characteristics and affordances of online spaces (e.g., anonymity) where user interaction and past user behaviour are strong indicators of a potentially concerning mental state of online users, particularly youth, that must be identified. In this study, employing both qualitative and quantitative methods, we address this gap by investigating the impact of augmenting text based suicidal ideation detection models

with contextual cues based on historical tweeting behavior and social media engagement.

To this end, we perform a survey of current methods and offset their limitations by developing a novel ensemble learning pipeline which lays the foundation for employing temporal and social interaction features to aid in identifying the subtleties and nuances in potentially concerning suicidal tweets. We perform an exploratory analysis and go over the limitations and ethical considerations pertaining to our proposed approach to pave the way for future work. We aim to provide valuable information to clinicians, health practitioners and policy makers to draw inferences regarding the behaviour exhibited by users suffering from suicide ideation.

## 10 ACKNOWLEDGEMENTS

Rajiv Ratn Shah is partly supported by the Infosys Center for AI, IIIT Delhi.

## REFERENCES


[1] Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976* (2016).

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 759–760.

[3] Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1373–1378.

[4] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538* (2017).

[5] Steven Bird and Edward Loper. 2004. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 31.

[6] Louise Brådvik, Cecilia Mattisson, Mats Bogren, and Per Nettelbladt. 2008. Long-term suicide risk of depression in the Lundby cohort 1947–1997–severity and gender. *Acta Psychiatrica Scandinavica* 117, 3 (2008), 185–191.

[7] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*. ACM, 75–84.

[8] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J Bierut. 2016. A content analysis of depression-related tweets. *Computers in human behavior* 54 (2016), 351–357.

[9] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (Atlanta GA*.

[10] Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications* 73 (2016), 291–300.

[11] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[12] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research* 7, Jan (2006), 1–30.

[13] Jasper Friedrichs, Debanjan Mahata, and Shubham Gupta. 2018. InfyNLP at SMM4H task 2: stacked ensemble of shallow convolutional neural networks for identifying personal medication intake from Twitter. *arXiv preprint arXiv:1803.07718* (2018).

[14] Madelyn Gould, Patrick Jamieson, and Daniel Romer. 2003. Media contagion and suicide among the young. *American Behavioral Scientist* 46, 9 (2003), 1269–1284.

[15] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 855–864.

[16] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[17] Nina Jacob, Jonathan Scourfield, and Rhiannon Evans. 2014. Suicide prevention via the internet. *Crisis* (2014).

[18] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis* (2014).

[19] Abhinav Khattar, Karan Dabas, Kshitij Gupta, Shaan Chopra, and Ponnurangam Kumaraguru. 2018. White or Blue, the Whale gets its Vengeance: A Social Media Analysis of the Blue Whale Challenge. *arXiv preprint arXiv:1801.05588* (2018).

[20] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[21] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[22] Depeng Liang and Yongdong Zhang. 2016. AC-BLSTM: Asymmetric Convolutional Bidirectional LSTM Networks for Text Classification. *arXiv preprint arXiv:1611.01884* (2016).

[23] Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, et al. 2018. # phramacovigilance-Exploring Deep Learning Techniques for Identifying Mentions of Medication Intake from Twitter. *arXiv preprint arXiv:1805.06375* (2018).

[24] Debanjan Mahata, John R Talburt, and Vivek Kumar Singh. 2015. From chirps to whistles: discovering event-specific informative content from Twitter. In *Proceedings of the ACM web science conference*. ACM, 17.

[25] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

[26] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1088–1098.

[27] Bridianne O'Dea, Melinda R Achilles, Mark E Larsen, Philip J Batterham, Alison L Calear, and Helen Christensen. 2018. The rate of reply and nature of responses to suicide-related posts on Twitter. *Internet interventions* 13 (2018), 105–107.

[28] Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2, 2 (2015), 183–188.

[29] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[30] Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. 2018. Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection. *arXiv preprint arXiv:1804.03124* (2018).

[31] J Ross Quinlan et al. 1996. Bagging, boosting, and C4. 5. In *AAAI/IAAI, Vol. 1*. 725–730.

[32] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen OâĂŹConnor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics* 54 (2015), 202–212.

[33] Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and Learning Suicidal Ideation Connotations on Social Media with Deep Learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 167–175.

[34] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*. 91–98.

[35] Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. 2018. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. *Journal of medical Internet research* 20, 5 (2018), e168.

[36] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[37] David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2 (1992), 241–259.

[38] RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[39] Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics* 5 (2017), 295–307.

[40] Yang Yu, Xiaojun Wan, and Xinjie Zhou. 2016. User embedding for scholarly microblog recommendation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 449–453.

[41] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social media mining: an introduction*. Cambridge University Press.

[42] Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *arXiv preprint arXiv:1803.03662* (2018).

[43] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1513–1522.

[44] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639* (2016).